The AFRL-MITLL WMT15 System: There's More than One Way to Decode It!

Jeremy Gwinnup[†], **Timothy Anderson**, **Grant Erdmann**, **Katherine Young**[†], **Christina May**[†]
Air Force Research Laboratory
jeremy.gwinnup.ctr,timothy.anderson.20,
grant.erdmann,katherine.young.1.ctr,
christina.may.ctr@us.af.mil

Michaeel Kazi[‡], **Elizabeth Salesky**[‡], **Brian Thompson**[‡]
MIT Lincoln Laboratory
michaeel.kazi,elizabeth.salesky,
brian.thompson@ll.mit.edu

Abstract

This paper describes the AFRL-MITLL statistical MT system and the improvements that were developed during the WMT15 evaluation campaign. As part of these efforts we experimented with a number of extensions to the standard phrase-based model that improve performance on the Russian to English translation task creating three submission systems with different decoding strategies. Out of vocabulary words were addressed with named entity postprocessing.

1 Introduction

As part of the 2015 Workshop on Machine Translation (WMT15) shared translation task, the MITLL and AFRL human language technology teams participated in the Russian-English translation task. Our machine translation systems represent enhancements to both our systems from IWSLT2014 (Kazi et al., 2014) and WMT14 (Schwartz et al., 2014), the addition of hierarchical decoding systems (Hoang and Koehn, 2008), neural network joint models (Devlin et al., 2014) and the utilization of Drem (Erdmann and Gwinnup, 2015) during the system tuning process.

2 System Description

We submitted systems for the Russian-to-English machine translation shared task. In all submitted systems, we used either phrase-based or hierarchical variants of the `moses` decoder (Koehn et al., 2007). As in previous years, our submitted systems used only the constrained data supplied when training.

2.1 Data Usage

In training our Russian-English systems we utilized the following corpora to train translation and language models: Yandex[1], Commoncrawl (Smith et al., 2013), LDC Gigaword English v5 (Parker et al., 2011) and News Commentary. The Wiki-Names corpus was reserved to train named entity recognizers.

2.2 Data Preprocessing

As with our WMT14 submission systems, preprocessing to address issues with the training data was required to ensure optimal system performance. Unicode characters in the private use, control character(C0, C1, zero-width, non-breaking, joiner, directionality and paragraph markers), and unallocated ranges were removed. Punctuation normalization and tokenization using Moses preprocessing scripts were then applied before lowercasing the data. The Commoncrawl corpus was further processed to exclude wrong-language text and to normalize mixed-alphabet spellings.

2.3 Factored Data Generation

We generated a class-factored version of the parallel Russian-English training data by using `mkcls` to produce 600 word classes for each side of the data. The factored data was then used to create a factored translation model and an in-domain class language model (Brown et al., 1993) for the English portion.

2.4 Phrase and Rule Table Training

Phrase tables and rule tables were trained on the preprocessed data using scripts provided with the `moses` distribution. Both rule tables and phrase tables utilized Good-Turing discounting (Gale, 1995). Hierarchical lexicalized reordering mod-

[†] Distribution A: Public Release
This work is sponsored by the Air Force Research Laboratory under Air Force contract FA-8650-09-D-6939-029.

[‡]This work is sponsored by the Air Force Research Laboratory under Air Force contract FA-8721-05-C-0002.

[1] https://translate.yandex.ru/corpus?lang=en

els (Galley and Manning, 2008) were also trained for use in the phrase-based systems.

An additional phrase table was trained on the lemmatized forms of the Russian training data. These lemmatized forms were generated by the `mystem`[2] tool.

2.5 Language Model Training

The English data sources listed in Section 2.1 were used to train a very large 6-gram language model (BigLM15). The English portion of the parallel data was processed into class form as outlined in Section 2.3 to generate an in-domain 600 class language model. `kenlm` (Heafield, 2011) was used to train these 6-gram models. These models were then binarized and stored on local solid-state disks for each machine in our cluster to improve load time and reduce fileserver traffic.

2.6 Operation Sequence Models

Using both the Russian and English data generated in Section 2.3, we trained order-5 Operation Sequence models (Durrani et al., 2011) for both the surface and class-factored forms of the data. These models were then used in our factored phrase-based system.

2.7 Neural Network Joint Models

NNJM (Devlin et al., 2014) models were used to rescore n-best lists. We trained these models on the alignments produced by `mgiza` (Gao and Vogel, 2008) over the parallel text. As in (Devlin et al., 2014), we trained four different models. The standard model is "source-to-target, left-to-right," which evaluates $p(t_i|T, S)$ with target window $T = (t_{i-1}, t_{i-2}, \ldots, t_{i-n})$ and $S = (s_{k-m}, \ldots s_k, \ldots, s_{k+m})$, where s_k is word-aligned to t_i. The four permutations of this are defined by (a) whether to count upwards from i, instead of downwards (this is left-to-right vs right-to-left), and (b) whether to swap the sources and targets entirely (source-to-target vs target-to-source).

We experimented with NNJM decoding (via a simple feature function in Moses). We achieved some benefit (+0.48) with this approach but rescoring a single NNJM source-to-target on 200-best lists produced better results in this case (+0.90). This was on a single system tuned on `newstest2013`, tested on `newstest2014` (baseline 29.07). In testing, 2-hidden layer rescoring models outperformed the 1-hidden layer decoding model.

Additionally, we experimented with vocabulary sizes determined by minimum count, trying 20 and 25. Using 20, our vocabulary was approximately 80,000 Russian words and 40,000 English; with 25, it was 70,000 and 34,000, respectively. We compared rescoring with a single, standard model (s2t, l2r) to rescoring with all directions with results listed in Table 1. We found that our max scores were better rescoring with all four models.

	Baseline	1 NNJM		4 NNJMs	
		20	25	20	25
max	27.71	27.90	28.05	27.90	28.07
mean	27.48	27.61	27.81	27.67	27.60

Table 1: NNJM Rescoring on `newstest2015`, optimizing on `newstest2014`.

2.8 Processing of Unknown Words

In our submission systems, we allowed words unknown to the decoder to be passed through to the translated output. The output of phrase-based and hierarchical submissions systems was then processed with permissive named entity lookup and selective transliteration. Our factored phrase-based system's output had named entity tagging and then permissive lookup and selective transliteration applied. Both techniques are described below. Score improvements in uncased BLEU are reported in Table 2. We see that application of permissive lookup and selective transliteration yielded an improvement of +0.48 BLEU versus a baseline system, while the application of named entity processing, permissive lookup and selective transliteration yielded a +0.57 BLEU gain.

2.8.1 Processing Unknown Words as Named Entities

The named entity post-process uses Russian-English pairs in the combined Wikinames and Wikititles lists (the wiki pairs list) and a transliteration-mined list to replace unknown words with English equivalents. We began by stemming each list to remove Russian noun and adjective endings. To the wiki pairs list, we added additional pairs yielded by replacing word-internal punctuation marks in existing wiki pairs with spaces. We used `giza++` (Och and Ney,

[2] `https://tech.yandex.ru/mystem`

2003) to align Russian-English phrases from the wiki list. We then used these alignments to start a generated list of pairs with only one Russian word and one English word in a pair. Of the aligned pairs, we only included pairs that were aligned with one another three or more times. Only one-to-one alignments would count toward the three alignment rule. We also removed entries where the English word in the pair occurred in a list of stop-words as well as where the English word consisted of only digits. To the generated list, we also added pairs directly from the wiki list with both single Russian words and single English words. Finally, we also added the highest quality pairs from the transliteration-mined list described below.

Upon encountering a single word without word-internal punctuation, the system first searches through the generated list, and returns a list of found guesses. If no items are found in the generated list, the wiki list is then searched. If still no guesses are found, then the transliteration-mined list is searched. The same process occurs for a word containing word-internal punctuation, but after a failed iteration of the search process, the punctuation is replaced with a space and the wiki lists are searched. Finally if that iteration fails, then the search process occurs on each individual word and a concatenation of English definitions is added to the guess list for every possible combination of guesses for each component word. An English language model is used to choose among the guesses.

2.8.2 Permissive Lookup and Selective Transliteration of Unknown Words

The second step focuses on selective transliteration of NE among the OOV words. We hypothesize that retaining transliterated forms of NE will improve readability, even if the output is not a direct match to the English reference.

The named entity pairs were harvested from the Common Crawl using NE tagging and rule-based transliteration matching. We used `mystem` to tag Russian NE words, and then compared them to capitalized English words in the parallel sentence, using an edit distance based on the typical sound values for the Cyrillic and Latin letters. We also searched for transliteration matches for capitalized words that were not tagged by Mystem, excluding sentence-initial words. Transliteration matches with a zero edit distance were added to the list of NE pairs used in the initial named entity processing, above. In this second step, we expand the list to include pairs with greater edit distance when they are validated by repeat occurrence.

If named entity processing as outlined in Section 2.8.1 was not applied, the named entity pairs list was expanded by adding the Wikinames and Wikititles lists.

After permissive named entity lookup, we attempt to distinguish NE from common words on the basis of capitalization in the Russian source file. Capitalized words that do not begin a sentence are assumed to be NE, and are transliterated. Lowercased words, and capitalized words that begin a sentence, are assumed to be common words and are dropped from the output.

3 Results

We submitted three systems for evaluation, each employing a different decoding strategy: Traditional phrased-based, Hierarchical, and Factored phrased-based. Each system is described below. Automatically scored results reported in BLEU (Papineni et al., 2002) for our submission systems can be found in Table 3.

Finally, as part of WMT15, the results of our submission systems listed in Tables 3 were ranked by monolingual human judges against the machine translation output of other WMT15 participants. These judgements are reported in WMT (2015).

3.1 Phrased-Based

We used a standard phrase based approach, using lowercased data. The lemma-based phrase table described in Section 2.4 was used as a back-off phrase table. We trained a hierarchical lexicalized reordering model, and used two separate class based (factored) language models; one using 600 classes on the in-domain target-side parallel data, and the other using the LDC Gigaword-English v5 NYT corpus. N-best lists from moses were rescored with 4-way NNJMs, and the system weights were tuned with PRO (Hopkins and May, 2011). Selective transliteration as described in Section 2.8.2 was then applied to the decoder output.

3.2 Hierarchical

New for this year, we trained a hierarchical system using the same parallel data as our phrase-based systems. The rule table was created as outlined in Section 2.4 and then filtered to only con-

System	Process Applied	baseline BLEU	postproc BLEU	Δ BLEU
phrase-based	PermLookup + SelTranslit	27.72	28.20	+0.48
hiero	PermLookup + SelTranslit	27.43	27.91	+0.48
pb-factored	NEProc + PermLookup+ SelTranslit	27.18	27.75	+0.57

Table 2: NEProc + SelTranslit Post-processing improvement measured in uncased BLEU

tain rules relating to the Russian content of the newstest test set for years 2012-2015. This filtering was performed in order to reduce the size of the rule table for both system memory requirements and expediency. The incremental-search algorithm (Heafield et al., 2013) and BigLM15 were used to decode the dev (newstest2014) and test (newstest2015) data. Drem was employed to optimize this system with Expected BLEU and Expected Meteor(Denkowski and Lavie, 2014) metrics. Finally, selective transliteration was employed as described in Section 2.8.2.

3.3 Factored Phrase-Based

For our last system, we used a factored phrase based approach (Koehn and Hoang, 2007) where the surface form of the training data was augmented with word classes. These classes were generated on the parallel training data outlined in Section 2.4 using mkcls to group the words into 600 classes for both English and Russian portions of the parallel training corpus. A phrase table and hierarchical reordering model was then trained using the moses training process on both the surface form and the class factor. Order-5 operation sequence models were separately trained on the surface forms and the class factors. An order-6 class-factor LM (Shen et al., 2006) was also trained on the English portion of the parallel training data to supplement the use of BigLM15. NNJMs as outlined in Section 2.7 were used to rescore the n-best lists from the decode. Following this rescoring, Drem was employed to optimize feature weights using the Expected Corpus BLEU metric (Smith and Eisner, 2006). After optimization and decoding of the test set, remaining unknown words were processed as described in Sections 2.8.1 and 2.8.2.

4 Discussion

Our three submitted systems all scored similarly against the official test set. Manual examination of our systems' output shows that there are significant differences in sentence structure and content.

System	Cased BLEU	Uncased BLEU
phrase-based	27.0	28.2
hiero	26.7	27.9
pb-factored	26.4	27.8

Table 3: MT Submission Systems decoding newstest2015

We scored one system output against another (as reference) with mteval13a.pl in both directions as BLEU scores are not symmetric. Results are listed in Table 4. Interestingly, the factored phrase based and hierarchical systems were more similar to each other than to the traditional phrase-based system. This suggests that the addition of class factors serves a similar function to the use of hierarchical decoding.

Test	Ref	BLEU
PB	Hiero	57.18
PBFac	Hiero	76.34
Hiero	PB	57.09
PBFac	PB	60.54
PB	PBFac	60.47
Hiero	PBFac	70.18

Table 4: Submission system similarity measured in uncased BLEU

It will be interesting to see the results of human evaluation on three markedly different systems.

5 Conclusion

In this paper, we present data preparation and processing techniques for our Russian-English submissions to the 2015 Workshop on Machine Translation (WMT15) shared translation task. Our submissions examine three different decoding strategies and the effectiveness of sophisticated handling of unknown words. While scoring similarly, each system produced markedly different output.

References

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. Proceedings of the ACL, Long Papers, Baltimore, MD, USA.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pages 1045–1054, Portland, Oregon, June.

Grant Erdmann and Jeremy Gwinnup. 2015. Drem: The AFRL submission to the WMT15 tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT'15)*, Lisbon, Portugal, September.

William A. Gale. 1995. Good-turing smoothing without tears. *Journal of Quantitative Linguistics*, 2.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June.

Kenneth Heafield, Philipp Koehn, and Alon Lavie. 2013. Grouping language model boundary words to speed k-best extraction from hypergraphs. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 958–968, Atlanta, Georgia, USA, June.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Hieu Hoang and Philipp Koehn. 2008. Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 58–65.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 1352–1362, Edinburgh, Scotland, U.K.

Michaeel Kazi, Elizabeth Salesky, Brian Thompson, Jessica Ray, Michael Coury, Tim Shen, Wade Anderson, Grant Erdmann, Jeremy Gwinnup, Katherine Young, Brian Ore, and Michael Hutt. 2014. The MIT-LL/AFRL IWSLT-2014 MT system. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT'14)*, Lake Tahoe, California, December.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 311–318, Philadelphia, Pennsylvania, July.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition LDC2011T07. *Philadelphia: Linguistic Data Consortium.*

Lane Schwartz, Timothy Anderson, Jeremy Gwinnup, and Katherine Young. 2014. Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT'14)*, pages 186–194, Baltimore, Maryland, USA, June.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 11 Jun 2015. Originator reference number RH-15-114103. Case number 88ABW-2015-2973.

Wade Shen, Richard Zens, Nicola Bertoldi, and Marcello Federico. 2006. The JHU workshop 2006 IWSLT system. In *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, November.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 787 – 794, Sydney, Australia.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 1374–1383, Sofia, Bulgaria, August.

WMT. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT '15)*, Lisbon, Portugal, September.